Data Analysis and Visualization in R (IN2339)

Exercise Session 2 - Data Wrangling

Daniela Klaproth-Andrade, Julien Gagneur

In this exercise session, we are analyzing an adapted version of a data set for book ratings, which contains 278,858 users (anonymized but with demographic information) providing 1,149,780 ratings about 271,379 books. We provide three different files containing information on the users, books and ratings. [https://www.kaggle.com/ruchi798/bookcrossing-dataset]

Section 00 - Getting ready

1. Make sure you have already installed and loaded the libraries `data.table` and `magrittr` by running the following commands:

```
install.packages("data.table")
install.packages("magrittr")
library(data.table)
library(magrittr)
```

Section 01 - Reading and cleaning up data

1. Load the three given datasets as `data.tables` and name them as `users_dt`, `books_dt` and `ratings_dt` accordingly. *Hint:* `fread()`

2. Check the classes of `users_dt`, `ratings_dt` and `books_dt`. Confirm that these are indeed a `data.table`.

3. Check the column names and classes of the `users_dt` data table and change the type of the `Age` column in `users_dt` to numeric.

4. Produce a summary of the variables in `books_dt`.

5. Return the first 5 and last 5 observations of the table `ratings_dt`.

6. Replace all the `-` in column names by underscores `_` in all three data tables. For example, `Book-Title` should be renamed to `Book_Title`. *Hint:* You can use the function `gsub()` that replaces pattern in a character string by a defined replacement. For example, for replacing `R` by `DataViz` in the following sentence `s` we use:

```
s <- 'R is fun'
gsub('R', 'DataViz', s)
```

```
## [1] "DataViz is fun"
```

7. Delete the columns `Image-URL-S`, `Image-URL-M` and `Image-URL-L` in the table `books_dt`.

8. What is the first year of publication? What is the last one?

9. Remove all the books published before 1900 and later than 2019 from `books_dt`.

Section 02 - Data Exploration

1. How many different authors are included in the table `books_dt`?

2. How many different authors are included for each year of publication between 2000 and 2010 in `books_dt`?

3. In how many observations is the age information missing in the users table `users_dt`?

4. Have a look at all locations from teenager users the table `users_dt`.

5. What is the maximum rating value in the ratings table?

6. What is the most common rating value larger than 0?

7. Which are the book identifiers (ISBN) with the highest ratings?

8. Sort the ratings table according to the rating value of each book in descending order. *Hint*: `order()`

9. Create a new column `Country` in the table `users_dt` for the name of the country of each user. For instance, from the location `cologne, nrw, germany`, we can assume the user comes from `Germany`. *Hint:* `tstrsplit()`

10. How many different countries are contained in the table `users_dt`?

11. What is the average age of the users in `users_dt`? What is the average age for users in NYC, Stockton and Moscow? *Hint:* use `by:=` and `i` for row filtering

Section 03 - Manipulating data tables

1. Add a new column called `High_Rating` to the data table `ratings_dt`. The column has an integer 1 for all observations with a rating value higher than 7.

2. How many observations are considered to be a high ranking? What is the proportion of high ranked observations among all observations?

3. Set the book identifier the key of the data table `books_dt`. What happened to the order of the data table? *Hint*: `setkey()`

4. Which users did not give any rating to any book? Filter these users out from `users_dt`. *Hint*: There's no need to merge `users_dt` with `ratings_dt`, we are simply interested in the users that are not in `ratings_dt`.

5. What is the most common age of users who rated at least one book?

6. On average, how many books did a user rate?

7. What is the title of the first published book with the highest ranking?

8. In which year was a book with the largest number of ratings last published?

9. Add to the table `ratings_dt` the highest ranking that each book received as a new column called `Max_Book_Ranking`.

10. Subset the merged ratings table to contain only books written by the following authors:

```
authors <- c("Agatha Christie", "William Shakespeare", "Stephen King",
             "Ann M. Martin", "Carolyn Keene", "Francine Pascal",
             "Isaac Asimov", "Nora Roberts", "Barbara Cartland", "Charles Dickens")
```

How many ratings has each author? What is their max and average ranking?